

# RELAXATION METHODS FOR MIXED-INTEGER OPTIMAL CONTROL OF PARTIAL DIFFERENTIAL EQUATIONS

FALK M. HANTE AND SEBASTIAN SAGER

**ABSTRACT.** We consider integer-restricted control of systems governed by abstract semilinear evolution equations. This includes the problem of optimal control design for certain distributed parameter systems endowed with multiple actuators, where the task is to minimize costs associated with the dynamics of the system by choosing, for each instant in time, one of the actuators together with ordinary controls. We consider relaxation techniques that are already used successfully for mixed-integer optimal control of ordinary differential equations. Our analysis yields sufficient conditions such that the solution of the relaxed problem can be approximated with arbitrary precision by a solution satisfying the integer restrictions. The results are obtained by semigroup theory methods. The approach is constructive and gives rise to a numerical method. We supplement the analysis with numerical experiments.

## 1. INTRODUCTION AND PROBLEM FORMULATION

The factoring of decision processes interacting with continuous evolution plays an important role in model-based optimization for many applications. For example when laying out a chemical reactor, among determining continuous variables such as modeling inlet and outlet concentrations for maximizing extraction rates, the number of reaction columns or restrictions to standard equipment sizes versus its costs become key considerations. Such mixed-integer optimal control problems are therefore studied in different communities with different approaches. Most of these approaches address problems that are governed by systems of ordinary differential equations in Euclidean spaces, see [17] for a survey on this topic.

Total discretization of the underlying system obviously leads to typically large mixed-integer nonlinear programs. Hence, relaxation techniques have become an integral part of efficient mixed-integer optimal control algorithms, either in the context of branch-and-bound type methods or, more directly, by means of nonlinear optimal control methods combined with suitable rounding strategies. An important result is that the solution of the relaxed problem can be approximated with arbitrary precision by a solution fulfilling the integer requirements [16].

In this paper, we extend such relaxation techniques to problems that are governed by certain systems of partial differential equations. Motivating applications are for example to switch between reductive and oxidative conditions in order to maximize the performance in a monolithic catalyst [22], port switching in chromatographic separation processes [5, 11], or to optimize switching control within photochemical reactions [20]. Our problem setting also includes the switching control design in the sense that systems are equipped with multiple actuators and the optimizer has to choose one of these together with ordinary controls for each instant in time.

---

*Date:* February 24, 2012.

F.M. Hante is with Mathematics Center of Heidelberg (MATCH), and Interdisciplinary Center of Scientific Computing (IWR), University of Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany. [falk.hante@iwr.uni-heidelberg.de](mailto:falk.hante@iwr.uni-heidelberg.de).

S. Sager is with Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany. [sebastian@sager1.de](mailto:sebastian@sager1.de).

Concerning systems involving partial differential equations, such switching control design has already been studied using several techniques: In [14, Chapter 8] optimal switching controls are constructed for systems governed by abstract semilinear evolution equations by combining ideas from dynamic programming and approximations of the value function using viscosity solutions of the Hamilton-Jacobi-Bellman equations. Switching boundary control for linear transport equations using switching time sensitivities has been studied in [8]. Exemplary for the heat equation and based on variational methods, the controllability in case of switching among several actuators has been considered in [23] and null-controllability for the one-dimensional wave equation with switching boundary control has been considered in [7]. Based on linear quadratic regulator optimal control techniques and enumeration of the integer values for a fixed time discretization, optimal switching control of abstract linear systems has been considered in [10].

Our approach is complementary to the above, as we break the computationally very expensive combinatorial complexity of the problem by relaxation. This comes at the downside of providing only a suboptimal solution but, as we will see, with arbitrary small integer-optimality gap, depending on discretization.

We will be concerned with the following problem of mixed-integer optimal control: Minimize a cost functional

$$J = \phi(y(t_f)) + \int_0^{t_f} L(y(t), u(t)) dt \quad (1)$$

over trajectories  $y: [0, t_f] \rightarrow X$  and control functions  $[u, v]: [0, t_f] \rightarrow U \times V$  subject to the constraints that  $y$  is a mild solution of the operator differential equation

$$\begin{cases} \dot{y}(t) = Ay(t) + f(t, y(t), u(t), v(t)), & t \in (0, t_f] \\ y(0) = y_0 \in X \end{cases} \quad (2)$$

and that the control functions satisfy

$$u(t) \in U_{\text{ad}} \subset U, \quad v(t) \in V_{\text{ad}} \subset V, \quad t \in [0, t_f] \quad (3)$$

where  $X$ ,  $U$  and  $V$  are Banach spaces,  $A: D(A) \rightarrow X$  is the infinitesimal generator of a strongly continuous semigroup  $\{T(t)\}_{t \geq 0}$  on  $X$ ,  $t_f \geq 0$  is a fixed real number,  $f: [0, t_f] \times X \times U \times V \rightarrow X$ ,  $\phi: X \rightarrow \mathbb{R}$  and  $L: X \times U \rightarrow \mathbb{R}$  are given functions,  $U_{\text{ad}}$  is some subset of  $U$  and  $V_{\text{ad}}$  is a *finite* subset of  $V$ .

We will refer to the above infinite-dimensional dynamic optimization problem as *mixed-integer optimal control problem*, short (MIP), and to the control function  $[u, v]$  as a *mixed-integer control*. This accounts for the fact that we do not impose restrictions on the set  $U_{\text{ad}} \subset U$  while we can always identify the finite set  $V_{\text{ad}} \subset V$  of the feasible control values for  $v$  with a finite number of integers

$$V_{\text{ad}} = \{v^1, \dots, v^N\} \simeq \{1, \dots, N\}. \quad (4)$$

Moreover, the operator differential equation (2) is an abstract representation of certain initial-boundary value problems governed by linear and semilinear partial differential equations, see, e. g., [15].

The existence of an optimal solution of the problem (MIP) depends, inter alia, on the spaces where we seek  $u: [0, t_f] \rightarrow U \times V$  and  $v: [0, t_f] \rightarrow V$ . Common choices are, for  $u, v$  respectively,  $L^2(0, t_f; U)$ , piecewise  $C^k(0, t_f; U)$  or (piecewise)  $H^k(0, t_f; U)$  for  $u$  and  $L^\infty(0, t_f; V)$  or the space of piecewise constant functions with values in  $V$  for  $v$ . We defer these considerations by assuming later that there exists an optimal solution of a related (to a certain extent convexified and relaxed) optimal control problem and present sufficient conditions guaranteeing that the solution of the relaxed problem can be approximated with arbitrary precision by a solution satisfying the integer restrictions.

This relaxation method becomes most easily evident from writing problem (MIP) using a differential inclusion, that is, minimize (1) subject to the constraints that  $y$  is a solution of

$$\begin{cases} \dot{y}(t) \in Ay(t) + \{f(t, y(t), u(t), v^i) : v^i \in V_{\text{ad}}\}, & t \in (0, t_f] \\ y(0) = y_0. \end{cases} \quad (5)$$

and  $u$  satisfies

$$u(t) \in U_{\text{ad}}, \quad t \in [0, t_f].$$

It is well known, that under certain technical assumptions, the solution set of (5) is dense in the solution set of the convexified differential inclusion

$$\begin{cases} \dot{y}(t) \in Ay(t) + \overline{\text{co}}\{f(t, y(t), u(t), v^i) : v^i \in V_{\text{ad}}\}, & t \in (0, t_f] \\ y(0) = y_0, \end{cases} \quad (6)$$

where  $\overline{\text{co}}$  denotes the closure of the convex hull. This is proved in [6] for the case when  $X$  is a separable Banach space and in [4] for non-separable Banach spaces. While these results rely on powerful selection theorems, our main contribution is a constructive proof based on discretization, giving rise to a numerical method at the prize of additional regularity assumptions.

We will see that the advantage of such a relaxation method is that the convexified problem, using a particular representation of (6), falls into the class of optimal control problems with partial differential equations without integer-restrictions. The already known theory, in particular concerning existence, uniqueness and regularity of optimal solutions as well as numerical considerations such as sensitivities, error analysis for finite element approximations, etc., can thus be carried over to the mixed-integer problem under consideration here. The disadvantage of this approach is that we target at a solution that is only suboptimal (though with arbitrary precision) and that switching costs, a standard regularization of mixed-integer problems to prevent chattering solutions, or additional combinatorial constraints can lead to larger optimality gaps. Nevertheless, we will show how a-priori bounds for such a gap can be obtained when constraints on the number of switches are incorporated.

The framework we use for the analysis here will be semigroup theory. Recall that for given  $y_0 \in X$  and given control functions  $u, v$ , the mild solution of the state equation (2) is given by a function  $y \in C(0, t_f; X)$  satisfying the variation of constants formula

$$y(t) = T(t)y_0 + \int_0^t T(t-s)f(s, y(s), u(s), v(s)) ds, \quad 0 \leq t \leq t_f \quad (7)$$

in the Lebesgue-Bochner sense. This abstract setting covers in particular the usual setup for weak solutions of linear parabolic partial differential equations with distributed control on reflexive Banach spaces where  $A$  arises from a time-invariant variational problem, see [1, Section 1.3].

Throughout the paper we denote by  $H_{\text{pw}}^1(0, t_f; X)$  the space of  $X$ -valued functions defined on the interval  $[0, t_f]$  and being piecewise once-weakly differentiable with a piecewise defined weak derivative that is square-integrable in the Lebesgue-Bochner sense. Consistently, we denote by  $C_{\text{pw}}^{0, \vartheta}(0, t_f; X)$  the space of  $X$ -valued functions defined on the interval  $[0, t_f]$  being piecewise Hölder-continuous with a Hölder-constant  $\vartheta$ . In both constructions, piecewise means that there exists a finite partition of the interval  $[0, t_f]$

$$0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_K < \tau_{K+1} = t_f \quad (8)$$

so that the function has the respective regularity on all intervals  $[\tau_k, \tau_{k+1})$ ,  $k = 0, \dots, K$ . We denote by  $\|\cdot\|_X$  the norm on  $X$  and by  $\|\cdot\|_{\mathcal{L}(X)}$  the operator norm

induced by  $\|\cdot\|_X$ . For simplicity of notation, we also define  $T(-t) = \text{Id}$  for all  $t > 0$ ,  $\text{Id}$  denoting the identity on  $X$ .

The paper is organized as follows. In Section 2, we present details of the relaxation method sketched above. In Section 3, the main result concerning an estimate of the approximation error is presented. In Section 4, we discuss extensions of the method to incorporate certain combinatorial constraints. In Section 5, we discuss applications for linear and semilinear equations and present numerical results for the heat equation with spatial scheduling of different actuators and a semilinear reaction-diffusion system with an on-off type control. In Section 6, we conclude with some additional remarks and point out open problems.

## 2. RELAXATION METHOD

Consider the following problem involving a particular representation of the convexified differential inclusion (6)

$$\left\{ \begin{array}{l} \text{minimize } J = \phi(y(t_f)) + \int_0^{t_f} L(y(t), u(t)) dt \quad \text{s. t.} \\ \dot{y}(t) = Ay(t) + \sum_{i=1}^N \alpha_i(t) f(t, y(t), u(t), v^i), \quad t \in (0, t_f] \\ y(0) = y_0 \\ u(t) \in U_{\text{ad}}, \quad t \in [0, t_f] \\ \alpha(t) = (\alpha_1(t), \dots, \alpha_N(t)) \in [0, 1]^N, \quad t \in [0, t_f] \\ \sum_{i=1}^N \alpha_i(t) = 1, \quad t \in [0, t_f]. \end{array} \right. \quad (9)$$

Observe that the control functions  $\alpha_i$  take values on the full interval  $[0, 1]$ , but that any optimal solution  $[u^*, \alpha^*]$  of (9) yields an optimal mixed-integer solution

$$[u^*, v^*] := [u^*, \sum_{i=1}^N \alpha_i^* v^i] \quad (10)$$

of problem (MIP) if  $\alpha^*(t) \in \{0, 1\}^N$  for almost every  $t \in (0, t_f)$ .

However, it is not very difficult to construct examples where  $\alpha^*(t) \in (0, 1)^N$  for  $t$  on some interval of positive measure. It is only in some special cases where optimality of  $\alpha^*(t) \in [0, 1]^N$  implies that  $\alpha^*$  takes only values on the boundary of its feasible set. For examples where this property, known as the bang-bang principle, can be verified in the context of partial differential equations, see [21, Section 3.2.4] and the references therein.

Therefore, under the main hypothesis

(H<sub>0</sub>) Problem (9) has an optimal solution  $[u^*, \alpha^*]$

we propose an iterative procedure in Algorithm 1 below to obtain from  $\alpha^*$  a control taking only values in  $\{0, 1\}^N$  and show in Theorem 1 under additional technical assumptions that the solution of problem (MIP) can be approximated by this procedure with arbitrary precision.

**Algorithm 1.**

- 1: Choose a time discretization grid  $\mathcal{G}^0 = \{0 = t_0^0 < t_1^0 < \dots < t_{n_0}^0 = t_f\}$  and some termination tolerance  $\varepsilon > 0$ . Set  $k = 0$ .
- 2: Find a relaxed optimal control  $[u^*, \alpha^*]: [0, t_f] \rightarrow U \times [0, 1]^N$  of (9) with a corresponding trajectory  $y^*$  and set  $J^* = J(u^*, \alpha^*, y^*)$ .
- 3: LOOP
- 4: Using  $\mathcal{G}^k$ , define a function  $\omega^k = (\omega_1^k, \dots, \omega_N^k): [0, t_f] \rightarrow \{0, 1\}^N$  piecewise constantly by

$$\omega_j^k(t) = p_{j,i}^k, \quad t \in [t_i^k, t_{i+1}^k) \quad (11)$$

where

$$p_{j,i}^k = \begin{cases} 1 & \text{if } \hat{p}_{j,i}^k \geq \hat{p}_{l,i}^k \text{ for all } l \neq j \text{ and } j < l \text{ with } \hat{p}_{j,i}^k = \hat{p}_{l,i}^k \\ 0 & \text{else} \end{cases} \quad (12)$$

$$\hat{p}_{j,i}^k = \int_0^{t_{i+1}^k} \alpha_j(\tau) d\tau - \sum_{l=0}^{i-1} p_{j,l}^k (t_{l+1}^k - t_l^k).$$

- 5: Set  $J^k = \phi(y^k(t_f)) + \int_0^{t_f} L(y^k(t), u^*(t)) dt$  where  $y^k$  is the solution of

$$\begin{cases} \dot{y}(t) = Ay(t) + \sum_{i=1}^N \omega_i^k(t) f(t, y(t), u^*(t), v^i), & t \in (0, t_f] \\ y(0) = y_0 \in X. \end{cases} \quad (13)$$

- 6: If  $J^k < J^* + \varepsilon$  then STOP.
- 7: Choose  $\mathcal{G}^{k+1} = \{0 = t_0^{k+1} < t_1^{k+1} < \dots < t_{n^{k+1}}^{k+1} = t_f\}$  such that  $\mathcal{G}^k \subset \mathcal{G}^{k+1}$  and set  $k = k + 1$ .
- 8: END LOOP
- 9: Set  $v^*(t) = \sum_{i=1}^N \omega_i^k(t) v^i, t \in [0, t_f]$ .

□

**Theorem 1.** *In addition to the main hypothesis (H<sub>0</sub>), assume that the following assumptions hold true.*

(H<sub>1</sub>) *The functions  $\phi: X \rightarrow \mathbb{R}$  and  $L: X \times U \rightarrow \mathbb{R}$  are continuous. Moreover, the function  $f$  satisfies the Lipschitz-estimate*

$$\|f(s, y_1, u^*(s), v^i) - f(s, y_2, u^*(s), v^i)\|_X \leq L \|y_1 - y_2\|_X$$

*with a constant  $L$  uniformly for  $t \in [0, t_f]$  a. e.,  $i = 1, \dots, N$  and  $y_1 \in \mathcal{Y}_1 = \{y^*(t) : t \in [0, t_f]\}$ ,  $y_2 \in \mathcal{Y}_2 = \{y(t) : t \in [0, t_f], y \text{ solves (9) for some } \alpha \in L^\infty(0, t_f; [0, 1]^N)\}$ .*

(H<sub>2</sub>) *For all  $i = 1, \dots, N$  and  $t \in [0, t_f]$  the function*

$$s \mapsto T(t-s)f(s, y^*(s), u^*(s), v^i)$$

*is in  $H_{pw}^1(0, t_f; X)$  and there exists a positive constant  $C_i$  such that*

$$\left\| \frac{d}{ds} T(t-s)f(s, y^*(s), u^*(s), v^i) \right\|_X \leq C_i, \quad \text{for a. e. } 0 < s < t < t_f.$$

*Let  $C = \sum_{i=1}^N C_i$ .*

(H<sub>3</sub>) *For all  $i = 1, \dots, N$ , there exists a positive constant  $M_i$  such that*

$$\sup_{t \in [0, t_f]} \|f(t, y^*(t), u^*(t), v^i)\|_X \leq M_i.$$

*Let  $M = \sum_{i=1}^N M_i$ .*

Moreover, assume that the sequence  $\{\mathcal{G}^k\}_k$  is such that  $\{\Delta t^k\}_k$  with

$$\Delta t^k = \max_{i=1, \dots, n^k} \{t_i^k - t_{i-1}^k\} \quad (14)$$

is strictly monotonically decreasing. Then Algorithm 1 terminates in a finite number of steps with a feasible solution  $[u^*, v^*]$  of Problem (MIP) satisfying the estimate

$$|J^* - J(u^*, v^*)| < \varepsilon. \quad (15)$$

*Remark 1.* The proof of Theorem 1 in Section 3 in fact yields the estimate

$$\|y^*(t) - y^k(t)\|_X \leq \left( (M + Ct)e^{\bar{M}Lt} \right) (N - 1)\Delta t^k, \quad t \in [0, t_f], \quad (16)$$

where  $\bar{M} = \sup_{t \in [0, t_f]} \|T(t)\|_{\mathcal{L}(X)}$  and the constants  $M, C$  and  $L$  are given by hypothesis (H<sub>2</sub>)–(H<sub>3</sub>). This proves a linear dependency of the integer-control approximation error in terms of the chosen maximal control discretization mesh size  $\Delta t^k$ . Moreover, assuming that in addition to (H<sub>1</sub>), there exists a constant  $\eta$  such that

$$|\phi(y_1) - \phi(y_2)| \leq \eta \|y_1 - y_2\|_X, \quad y_1, y_2 \in X \quad (17)$$

and a function  $\xi \in L^1(0, t_f)$  such that

$$|L(y_1, u^*(t)) - L(y_2, u^*(t))| \leq \xi(t) \|y_1 - y_2\|_X, \quad y_1, y_2 \in X \quad (18)$$

then one obtains from (16) by standard estimates the following error bound

$$|J^* - J([u^*, v^*])| \leq \left( \eta C(t_f) + \int_0^{t_f} \xi(t) C(t) dt \right) (N - 1)\Delta t^k, \quad (19)$$

with  $C(t) = (M + Ct)e^{\bar{M}Lt}$ . This shows again a linear dependency of the error on the chosen maximal control discretization mesh size  $\Delta t^k$ .

In a similar fashion, the method can also deal with state constraints. Suppose that we wish to include a constraint of the form

$$G(y(t), t) \geq 0, \quad t \in [0, t_f] \quad (20)$$

in the mixed-integer optimal control problem (MIP). Including this constraint also in (9) and assuming continuity of the function  $G: X \times [0, t_f] \rightarrow \mathbb{R}$ , then (16) yields that the constraint violation of the integer solution satisfies

$$|G(y^*(t), t) - G(y^k(t), t)| \leq \varepsilon \quad (21)$$

for  $k$  sufficiently large. Moreover, if we assume that there exists a function  $\zeta \in L^\infty(0, t_f)$  such that

$$|G(y_1, t) - G(y_2, t)| \leq \zeta(t) \|y_1 - y_2\|_X, \quad y_1, y_2 \in X \quad (22)$$

then (16) yields that

$$|G(y^*(t), t) - G(y^k(t), t)| \leq \zeta(t) C(t) (N - 1)\Delta t^k \quad (23)$$

with  $C(t)$  as in (19). This shows again a similar linear dependency on  $\Delta t^k$ . The assumption (22) holds for example for functions  $G$  enforcing time-periodicity constraints

$$y(t_f) = y(0), \quad (24)$$

occurring, e.g., in chromatographic separation processes [5, 11].

*Remark 2.* Algorithm 1 requires in step 2: to solve a relaxed optimal control problem without integer constraints. We note that the conclusion of Theorem 1 remains true if we replace the main hypothesis (H<sub>0</sub>) by

(H'<sub>0</sub>) Problem (9) has a feasible solution  $[u^*, \alpha^*]$ .

and, accordingly, waive the optimality of  $[u^*, \alpha^*]$  stipulated in step 2: of Algorithm 1. An important case, for instance, is when the control functions  $[u^*, \alpha^*]$  found in step 2: satisfy only local optimality conditions. This can still be useful in order to provide bounds for the mixed integer-problem (MIP). Nevertheless, an approximation of a globally optimal solution of problem (MIP) can of course only be obtained when the problem in step 2: is solved to global optimality.

Moreover, the problem in step 2: can in practice only be solved numerically and may therefore require further discretization, e. g., using the method of lines by means of a finite-element semi-discretization in space. Suppose that such a discretization is well-posed in the sense that

$$\|y_h^*(t) - y^*(t)\|_X \leq C(h)\|y^*\|_X, \quad t \in [0, t_f] \quad (25)$$

for a constant  $C(h) \rightarrow 0$  as  $h \rightarrow 0$ , where  $y_h^*$  is the solution of the discretized problem parameterized by  $h$ . The conclusion of Theorem 1 then again remains true if the control functions  $[u^*, \alpha^*]$  are obtained by means of a sufficiently small  $h$ . We then have the estimate

$$\|y_h^*(t) - y^k(t)\|_X \leq \|y_h^*(t) - y^*(t)\|_X + \|y^*(t) - y^k(t)\|_X \quad (26)$$

and see, from (16) in Remark 1 and (25) that

$$\|y_h^*(t) - y^k(t)\|_X \rightarrow 0 \quad (27)$$

for  $(h, \Delta t^k) \rightarrow 0$ .

*Remark 3.* In order to solve the optimal control problem (9) numerically and in view of Remark 2, the problem may have to be (adaptively) discretized. In particular, direct or indirect numerical methods may be used. For an introduction to the basic concepts see, e. g., [9]. Depending on the method of choice for the time discretization of the control function  $u: [0, t_f] \rightarrow U$  it may in many cases be advantageous to discretize  $u$  and  $v$  on the same grid  $\mathcal{G}^k$  and to move step 2: into the main loop of the algorithm with

3': Find a relaxed optimal control  $[u^k, \alpha^k]: \mathcal{G}^k \rightarrow U \times [0, 1]^N$  of (9) with a corresponding trajectory  $y^*$  and set  $J^* = J(u^k, \alpha^k, y^*)$ .

along with another termination criterion

3'': If  $\omega^k(t) := \alpha^k(t) \in \{0, 1\}$  for  $t \in [0, t_f]$ , then STOP.

preceding step 4: Then the conclusion of Theorem 1 still holds if the optimal relaxed solution  $J^*$  is frozen after a finite number of iterations  $k_{\max}$  by setting  $y^* = y(u^{k_{\max}}, \alpha^{k_{\max}})$  and  $J^* = J(u^{k_{\max}}, \alpha^{k_{\max}}, y^*)$ . This is implemented in the software package MS MINTOC designed for solving mixed-integer optimal control problems with ordinary differential equations [18, 17].

Clearly, hypothesis (H<sub>2</sub>) of Theorem 1 imposes additional regularity assumptions on the linear operator  $A$  generating the semigroup  $\{T(t)\}_{t \geq 0}$ , the function  $f$ , but also on the admissible time regularities of the control functions  $u: [0, t_f] \rightarrow U$  and  $v: [0, t_f] \rightarrow V$  or  $\alpha: [0, t_f] \rightarrow [0, 1]^N$ , respectively. The main difficulty is that  $y^*$  as a solution of (7) with  $A$  unbounded may only be continuous and not absolutely continuous in time, hence not necessarily differentiable almost everywhere as this is always true when  $A = 0$  (with  $T(\cdot) = \text{Id}$ ) and  $X$  is a finite dimensional space. A little more can be said for linear systems

$$\dot{y}(t) = Ay(t) + f(t, u(t), v(t)), \quad y(0) = y_0 \quad (28)$$

when  $f$  is sufficiently smooth.

*Remark 4.* Consider the problem (MIP) with equation (2) replaced by (28), let  $\bar{M} = \sup_{t \in [0, t_f]} \|T(t)\|_{\mathcal{L}(X)}$  and suppose that the functions  $g^i: [0, t_f] \rightarrow X$  defined by  $g^i(t) = f(t, u^*(t), v^i)$ ,  $i = 1, \dots, N$ , satisfy the following conditions

(i)  $g^i(t) \in D(A)$  for a. e.  $t \in [0, t_f]$  and there exists constants  $\bar{L}^i$  such that

$$\operatorname{ess\,sup}_{t \in [0, t_f]} \|Ag^i(t)\|_X \leq \bar{L}^i. \quad (29)$$

(ii)  $g^i$  is differentiable for a. e.  $t \in [0, t_f]$  and there exist constants  $\bar{C}^i$  such that

$$\operatorname{ess\,sup}_{t \in [0, t_f]} \left\| \frac{d}{dt} g^i(t) \right\|_X \leq \bar{C}^i. \quad (30)$$

Then, because of condition (i) we get from the chain rule that

$$\frac{d}{ds} T(t-s)g(t) = T(t-s) \frac{d}{ds} g(t) - T(t-s)Ag(t) \quad (31)$$

for all  $i = 1, \dots, N$  and thus, by taking the norm, applying the triangular inequality and using the definition of the constants from above

$$\left\| \frac{d}{ds} T(t-s)f(s, u^*(s), v^i) \right\|_X \leq \bar{M} (\bar{C}^i + \bar{L}^i). \quad (32)$$

So hypothesis (H<sub>2</sub>) of Theorem 1 holds with  $C_i := \bar{M} (\bar{C}^i + \bar{L}^i)$ . The conditions (i) and (ii) are a natural extension of the differentiability assumptions imposed in [16, Corollary 6] for the case when  $A = 0$  and  $X = \mathbb{R}^n$ .

We further discuss hypothesis (H<sub>2</sub>) in Section 5 for linear and semilinear systems exemplary for cases when  $A$  is the generator of an analytic semigroup.

### 3. PROOF OF THEOREM 1

We prove the following result.

**Lemma 1.** *Let  $\varepsilon > 0$  and  $\bar{M} = \sup_{t \in [0, t_f]} \|T(t)\|_{\mathcal{L}(X)}$ . Suppose that  $[u^*, \alpha^*, y^*]$  is a feasible solution of the relaxed problem (9) and assume that the hypotheses (H<sub>1</sub>)–(H<sub>3</sub>) of Theorem 1 hold true. Let  $\omega = (\omega_1, \dots, \omega_N) \in L^\infty(0, t_f; [0, 1]^N)$  be such that*

$$\max_{i=1, \dots, N} \sup_{t \in [0, t_f]} \left| \int_0^t \alpha_i^*(\tau) - \omega_i(\tau) d\tau \right| \leq \varepsilon \quad (33)$$

and let  $y$  be the mild solution of (13) with  $\omega_i^k = \omega_i$ ,  $i = 1, \dots, N$ . Then

$$\|y^*(t) - y(t)\|_X \leq (M + Ct)e^{\bar{M}Lt} \varepsilon, \quad (34)$$

for  $t \in [0, t_f]$ .

*Proof.* Fix  $t \in [0, t_f]$  and set, for the sake of brevity,  $\delta(t) = \|y^*(t) - y(t)\|_X$ . Recalling (8), let  $\{\tau_0, \tau_1, \dots, \tau_{K+1}\}$  be the set of partition points of the functions  $s \mapsto T(t-s)f(s, y^*(s), u^*(s), v^i) \in H_{\text{pw}}^1(0, t_f; X)$ ,  $i = 1, \dots, N$ , assumed in (H<sub>2</sub>). From the definition of the mild solutions to (9) and (13), we have

$$\delta(t) = \left\| \sum_{i=1}^N \int_0^t T(t-s)f(s, y^*(s), v^i)\alpha_i^*(s) - T(t-s)f(s, y(s), v^i)\omega_i(s) ds \right\|_X.$$

Adding  $0 = T(t-s)f(s, y^*(s), v^i)\omega_i(s) - T(t-s)f(s, y^*(s), v^i)\omega_i(s)$  under the integral, applying the triangular inequality and rearranging terms this yields

$$\begin{aligned} \delta(t) &\leq \sum_{i=1}^N \left\| \int_0^t T(t-s)[f(s, y^*(s), v^i) - f(s, y(s), v^i)]\omega_i(s) ds \right\|_X \\ &\quad + \sum_{i=1}^N \left\| \sum_{k=0}^K \int_{\tau_k}^{\tau_{k+1}} T(t-s)f(s, y^*(s), v^i)[\alpha_i^*(s) - \omega_i(s)] ds \right\|_X. \end{aligned}$$



Now using integration by parts in the second part, we obtain

$$\begin{aligned} \delta(t) &\leq \sum_{i=1}^N \left\| \int_0^t T(t-s) [f(s, y^*(s), v^i) - f(s, y(s), v^i)] \omega_i(s) ds \right\|_X \\ &\quad + \sum_{i=1}^N \left\| \sum_{k=0}^K \left( T(t - \tau_{k+1}) f(\tau_{k+1}, y^*(\tau_{k+1}), v^i) \int_0^{\tau_{k+1}} \alpha_i^*(s) - \omega_i(s) ds \right. \right. \\ &\quad \left. \left. - T(t - \tau_k) f(\tau_k, y^*(\tau_k), v^i) \int_0^{\tau_k} \alpha_i^*(s) - \omega_i(s) ds \right. \right. \\ &\quad \left. \left. - \int_{\tau_k}^{\tau_{k+1}} \frac{d}{ds} (T(t-s) f(s, y^*(s), v^i)) \int_0^s \alpha_i^*(\vartheta) - \omega_i(\vartheta) d\vartheta ds \right) \right\|_X. \end{aligned}$$

Then by rearranging terms, noting that

$$\begin{aligned} &\sum_{k=0}^K \left( T(t - \tau_{k+1}) f(\tau_{k+1}, y^*(\tau_{k+1}), v^i) \int_0^{\tau_{k+1}} \alpha_i^*(s) - \omega_i(s) ds \right. \\ &\quad \left. - T(t - \tau_k) f(\tau_k, y^*(\tau_k), v^i) \int_0^{\tau_k} \alpha_i^*(s) - \omega_i(s) ds \right) \\ &= f(t, y^*(t), v^i) \int_0^t \alpha_i^*(s) - \omega_i(s) ds \end{aligned}$$

because of terms canceling out,  $T(t-t) = \text{Id}$  and  $\int_0^0 \alpha_i^*(\vartheta) - \omega_i(\vartheta) d\vartheta = 0$ , and by applying the triangular inequality this estimate simplifies to

$$\begin{aligned} \delta(t) &\leq \sum_{i=1}^N \int_0^t \|T(t-s)\|_{\mathcal{L}(X)} \|f(s, y^*(s), v^i) - f(s, y(s), v^i)\|_X |\omega_i(s)| ds \\ &\quad + \sum_{i=1}^N \|f(t, y^*(t), v^i)\|_X \left| \int_0^t \alpha_i^*(s) - \omega_i(s) ds \right| \\ &\quad + \sum_{i=1}^N \int_0^t \left\| \frac{d}{ds} (T(t-s) f(s, y^*(s), v^i)) \right\|_X \left| \int_0^s \alpha_i^*(\vartheta) - \omega_i(\vartheta) d\vartheta \right| ds. \end{aligned}$$

Then, by definition of the constant  $\bar{M}$ , the definition of the constants  $L$ ,  $C$  and  $M$  in hypotheses (H<sub>1</sub>)–(H<sub>3</sub>), the assumption (33) and the fact that  $\omega_i(t) \leq 1$ , this yields

$$\delta(t) \leq \bar{M}L \int_0^t \delta(s) ds + M\varepsilon + Ct\varepsilon.$$

Finally, using the Gronwall lemma and rearranging terms, we obtain the desired estimate

$$\delta(t) \leq \left( (M + Ct)e^{\bar{M}Lt} \right) \varepsilon.$$

□

Moreover, in order to prove Theorem 1, we use a recent result on integral approximations.

**Lemma 2.** (Theorem 5 of [16]) Let  $\alpha = (\alpha_1, \dots, \alpha_N): [0, t_f] \rightarrow [0, 1]^N$  be a measurable function satisfying  $\sum_{i=1}^N \alpha_i(t) = 1$  for all  $t \in [0, t_f]$ . Let a function  $\omega: [0, t_f] \rightarrow \{0, 1\}^N$  be defined piecewise constantly by

$$\omega_j(t) = p_{j,i}, \quad t \in [t_i, t_{i+1}) \quad (35)$$

where  $0 = t_0 < t_1 < \dots < t_n = t_f$ ,

$$p_{j,i} = \begin{cases} 1 & \text{if } \hat{p}_{j,i} \geq \hat{p}_{l,i} \text{ for all } l \neq j \text{ and } j < l \text{ with } \hat{p}_{j,i} = \hat{p}_{l,i} \\ 0 & \text{else} \end{cases} \quad (36)$$

$$\hat{p}_{j,i} = \int_0^{t_{i+1}} \alpha_j(\tau) d\tau - \sum_{l=0}^{i-1} p_{j,l}(t_{l+1} - t_l).$$

Then it holds

$$(1) \quad \max_{i=1,\dots,N} \left| \int_0^t \alpha(\tau) - \omega_i(\tau) \right| \leq (N-1)\Delta t,$$

$$(2) \quad \sum_{i=1}^N \omega_i(t) = 1 \text{ for all } t \in [0, t_f],$$

where  $\Delta t = \max_{i=1,\dots,n} \{t_i - t_{i-1}\}$ .

With the above two results we are now in the position to prove Theorem 1.

*Proof of Theorem 1.* Let the assumptions of Theorem 1 hold true. First observe that any control  $[u^*, v^*]$  obtained by Algorithm 1 is feasible for the problem (MIP), because  $u^*$  is feasible by assumption and  $v^*(t) = \sum_{i=1}^N \omega_i^k(t) v^i$ ,  $t \in [0, t_f]$ , by construction in step 4:, where  $\omega_i^k(t) \in \{0, 1\}$  for all  $i, k$  and  $t \in [0, t_f]$  as seen from (11) and (12). Then suppose that the main loop in Algorithm 1 terminates in step 6:. Then, the termination criterion  $J_k^* < J^* + \varepsilon$  and step 9:, recalling the bijection (10), implies that  $|J^* - J(u^*, v^*)| < \varepsilon$ . Next suppose that Algorithm 1 loops infinitely many times, that is,

$$|J_k^* - J^*| \geq \varepsilon, \text{ for all } k = 1, 2, \dots \quad (37)$$

and consider  $J^k = |J_k^* - J^*|$  as a sequence of  $k$ . From Lemma 2 with  $\alpha = \alpha^*$ ,  $\omega = \omega^k$  and  $\Delta t = \Delta t_k$  with  $\Delta t_k$  from (14), we get that

$$\max_{i=1,\dots,N} \left| \int_0^t \alpha^*(\tau) - \omega_i^k(\tau) d\tau \right| \leq (N-1)\Delta t^k, \quad t \in [0, t_f]. \quad (38)$$

Then Lemma 1 used with  $\omega = \omega^k$  and  $y = y^k$  implies that

$$\|y^*(t) - y^k(t)\|_X \leq \left( (M + Ct)e^{\bar{M}Lt} \right) (N-1)\Delta t^k, \quad t \in [0, t_f]. \quad (39)$$

The assumption that  $\Delta t^k$  is a strictly monotonically decreasing sequence together with the continuity assumptions on  $\phi$  and  $L$  due to (H<sub>1</sub>) then implies that  $J^k = |J_k^* - J^*|$ , because

$$|J_k^* - J^*| = \left| \phi(y_k(t_f)) - \phi(y^*(t_f)) + \int_0^{t_f} L(y_k(t), u^*(t)) - L(y^*(t), u^*(t)) dt \right|,$$

is also a strictly monotonically decreasing sequence. This contradicts (37) and completes the proof.  $\square$

#### 4. COMBINATORIAL CONSTRAINTS

Suppose we wish to include combinatorial constraints of the form

$$\#_{v^i \curvearrowright v^j}(v) \leq K^{i,j}, \quad i \in I, \quad j \in J \quad (40)$$

into the mixed-integer optimal control problem (MIP) given by (1)–(3), where  $\#_{v^i \curvearrowright v^j}(v)$  denotes the number of switches of the control function  $v: [0, t_f] \rightarrow V_{\text{ad}}$  from value  $v^i$  to value  $v^j$ ,  $K^{i,j}$  are given, non-negative constants and  $I, J \subset \{1, \dots, N\}$ .

Note that the relaxation method considered in Section 2 typically satisfies

$$\#_{v^i \curvearrowright v^j}(v) \rightarrow \infty$$

for some  $i, j \in \{1, \dots, N\}$  as we let  $\varepsilon \rightarrow 0$ , so eventually violating (40) for small  $\varepsilon$ . Therefore, along the lines of [19], we propose a modification of Algorithm 1, where we replace steps 4: and 6: by

4': Using  $\mathcal{G}^k$ , define the function  $\omega^k = (\omega_1^k, \dots, \omega_N^k): [0, t_f] \rightarrow \{0, 1\}^N$  piecewise constantly by

$$\omega_j^k(t) = p_{j,i}^{k,*}, \quad t \in [t_i^k, t_{i+1}^k) \quad (41)$$

where  $p_{j,i}^{k,*}$  is given by the solution of the min-max problem

$$\left\{ \begin{array}{l} \min_{p^k} J_{\text{sub}}(p^k) = \max_{i=1, \dots, N} \max_{r=1, \dots, n^k} \left| \sum_{l=1}^r (q_{i,l}^k - p_{i,l}^k) \Delta t_l^k \right| \\ \text{subject to} \\ \sum_{r=1}^{n^k} |p_{i,r}^k - p_{j,r+1}^k| \leq K^{i,j}, \quad i \in I, j \in J \\ \sum_{i=1}^N p_{i,r} = 1, \quad r = 1, \dots, n^k \\ p_{i,r} \in \{0, 1\}, \quad i = 1, \dots, N, r = 1, \dots, n^k \end{array} \right. \quad (42)$$

with  $\Delta t_l^k = t_{l+1}^k - t_l^k$ ,  $l = 1, \dots, n^k$ , and

$$q_{i,l}^k = \frac{1}{\Delta t_l^k} \int_{t_l}^{t_{l+1}} \alpha_i^*(t) dt, \quad i = 1, \dots, N, l = 1, \dots, n^k. \quad (43)$$

6': If  $|J_{\text{sub}}(p^{k,*}) - J_{\text{sub}}(p^{k-1,*})| < \varepsilon$  or  $k > k_{\max}$  then STOP.

The min-max problem (42) can be written as a standard mixed-integer linear problem (MILP) using slack variables and can be computed efficiently [19]. We then have the following result.

**Theorem 2.** *Suppose that the hypotheses of Theorem 1 hold true. Let  $\bar{M} = \sup_{t \in [0, t_f]} \|T(t)\|_{\mathcal{L}(X)}$  and  $C(t) = (M + Ct)e^{\bar{M}Lt}$ , the constants  $M, C$  and  $L$  given by (H<sub>2</sub>)–(H<sub>3</sub>). Then the Algorithm 1 with steps 4: and 6: replaced with 4': and 6':, respectively, terminates for every  $\varepsilon > 0$  and  $k_{\max} \geq 0$  with a feasible solution  $[u^*, v^*]$  of problem (MIP) satisfying the combinatorial constraints (40) and the estimate*

$$\|y^*(t) - y^k(t)\|_X \leq C(t)(J_{\text{sub}}(p^{k,*}) + \delta) \quad (44)$$

for some  $0 \leq \delta \leq \max_{l=1, \dots, n^k} \Delta t_l^k$ .

If, in addition to (H<sub>1</sub>), there exists a constant  $\eta$  such that

$$|\phi(y_1) - \phi(y_2)| \leq \eta \|y_1 - y_2\|_X, \quad y_1, y_2 \in X \quad (45)$$

and a function  $\xi \in L^1(0, t_f)$  such that

$$|L(y_1, u^*(t)) - L(y_2, u^*(t))| \leq \xi(t) \|y_1 - y_2\|_X, \quad y_1, y_2 \in X \quad (46)$$

then the following estimate holds

$$|J^* - J([u^*, v^*])| \leq \left( \eta C(t_f) + \int_0^{t_f} \xi(t) C(t) dt \right) (J_{\text{sub}}(p^{k,*}) + \delta) \quad (47)$$

with the same  $\delta$  as in (44).

*Proof.* Algorithm 1 with steps 4: and 6: replaced with 4': and 6':, respectively, terminates after  $k$  steps,  $k \leq k_{\max}$ , with controls  $[u^*, v^*]$ . The control  $u^*$  is feasible for problem (MIP) by assumption. From (41) and the constraints in (42) we see that  $\omega_j^k(t) \in \{0, 1\}$  for all  $t \in [0, t_f]$  and thus, by step 9: in Algorithm 1 and recalling the bijection (10), the control  $v^*(t) = \sum_{i=1}^N \omega_i^k(t) v^i$ ,  $t \in [0, t_f]$  is feasible for problem (MIP). The constraints in (42) also ensure that the combinatorial constraints (40) are satisfied. Moreover, the cost function in (40) is defined as

$$J_{\text{sub}}(p^{k,*}) = \max_{i=1,\dots,N} \max_{r=1,\dots,n^k} \left| \sum_{l=1}^r (q_{i,l}^k - p_{i,l}^k) \Delta t_l^k \right|. \quad (48)$$

By definition of  $q_{i,l}^k$  in (43) and  $\omega_j^k(t)$  in (41) and rearranging terms, we get

$$J_{\text{sub}}(p^{k,*}) = \max_{i=1,\dots,N} \max_{r=1,\dots,n^k} \left| \int_0^{t_{r+1}} \alpha_i^*(t) - \omega_i^k(t) dt \right|. \quad (49)$$

Using that  $\alpha_i^*(t) \in [0, 1]$  and  $\omega_i^k(t) \in \{0, 1\}$  for all  $t \in [0, t_f]$ , this yields

$$J_{\text{sub}}(p^{k,*}) = \max_{i=1,\dots,N} \sup_{t \in [0, t_f]} \left| \int_0^t \alpha_i^*(\tau) - \omega_i^k(\tau) d\tau \right| - \delta \quad (50)$$

for some  $0 \leq \delta \leq \max_{l=1,\dots,n^k} \Delta t_l^k$ . Applying Lemma 1 with  $\varepsilon = J_{\text{sub}}(p^{k,*}) + \delta$  we get that

$$\|y^*(t) - y^k(t)\|_X \leq (M + Ct)e^{\bar{M}Lt}(J_{\text{sub}}(p^{k,*}) + \delta) \quad (51)$$

with the constants  $M, C$  and  $L$  given by (H<sub>2</sub>)–(H<sub>3</sub>). By definition of  $C(t)$ , this proves (44). The estimate (47) then follows from (44) using the definition of the cost function  $J$  in (1) and the Lipschitz-estimates (45) and (46). This completes the proof of Theorem 2.  $\square$

*Remark 5.* As already remarked in the case without combinatorial constraints, the method can also deal with state constraints such as (20). Assuming again existence of a function  $\zeta \in L^\infty(0, t_f)$  such that (22) holds true, (44) yields a bounded deviation of the feasible reference trajectory

$$|G(y^*(t), t) - G(y^k(t), t)| \leq \zeta(t)C(t)(J_{\text{sub}}(p^{k,*}) + \delta), \quad (52)$$

and hence a bound on the worst case constraint violation. Remark 2 and Remark 3 apply similarly.

## 5. EXAMPLES

In this section we discuss the hypothesis (H<sub>1</sub>)–(H<sub>3</sub>) of Theorem 1 exemplary for a linear and a semilinear control problem where  $A$  is the generator of an analytic semigroup and present numerical results for a test problem in each case.

**5.1. A linear parabolic equation with lumped controls.** Let  $\Omega$  be a domain in  $\mathbb{R}^n$  and  $f_i: \Omega \rightarrow \mathbb{R}$ ,  $i = 1, \dots, N$ , be fixed control profiles. Consider the internally controlled heat equation

$$\begin{cases} \frac{\partial y}{\partial t}(x, t) - \rho \sum_{j=1}^n \frac{\partial^2 y}{\partial x_j^2}(x, t) = f_{\sigma(t)}(x)u(t), & \text{in } Q \\ y(x, t) = 0, & \text{on } \Sigma \\ y(x, 0) = y_0(x), & \text{in } \Omega \end{cases} \quad (53)$$

where  $Q = \Omega \times (0, t_f)$ ,  $\Sigma = \partial\Omega \times (0, t_f)$  and  $\rho$  is a positive constant.

Suppose the control task is to minimize the cost function

$$J = \int_{\Omega} |y(t_f, x)|^2 dx + \lambda_1 \int_0^{t_f} \int_{\Omega} |y(t, x)|^2 dx dt + \lambda_2 \int_0^{t_f} |u(t)|^2 dt \quad (54)$$

where  $y$  is the weak solution of (53) by selecting  $u: [0, t_f] \rightarrow \mathbb{R}$  and a switching signal  $\sigma(\cdot): [0, t_f] \rightarrow \{1, \dots, N\}$  determining the control profile  $f_i$  applied at time  $t \in [0, t_f]$ .

In order to write the above problem in abstract form (2), we let  $X = L^2(\Omega)$ , set  $V = U = U_{\text{ad}} = \mathbb{R}$ ,  $V_{\text{ad}} = \{1, \dots, N\}$  and define  $f: [0, t_f] \times U \times V \rightarrow X$  by  $f(t, u, v)(x) := f_v(x)u$ ,  $\phi(y) = \|y\|_X^2$ ,  $L(y, u) = \lambda_1 \|y\|_X^2 + \lambda_2 |u|^2$  and define  $(A, D(A))$  as

$$\begin{aligned} D(A) &= H^2(\Omega) \cap H_0^1(\Omega) \\ (Ay)(x) &= \sum_{j=1}^n \frac{\partial^2 y}{\partial x_j^2}(x), \quad y \in D(A). \end{aligned} \quad (55)$$

It is well-known that  $(A, D(A))$  is the generator of a strongly continuous (analytic) semigroup of contractions  $\{T(t)\}_{t \geq 0}$  on  $X$ , see, e. g., [15].

Let  $[u^*, \alpha^*]$  be a feasible solution of the relaxed and convexified problem (9). Assume that these feasible controls  $[u^*, \alpha^*]$  satisfy

$$[u^*, \alpha^*] \in H_{\text{pw}}^1(0, T) \times L^\infty(0, T)^N \quad (56)$$

and that the fixed control profiles  $f_i$  satisfy

$$f_i \in D(A) \text{ for all } i = 1, \dots, N. \quad (57)$$

We now want to check if the assumptions (H<sub>1</sub>)–(H<sub>3</sub>) of Theorem 1 are satisfied.

The continuity assumptions (H<sub>1</sub>) for  $\phi$  and  $L$  and  $f$  clearly hold. From semigroup theory and the chain rule we get, using (57)

$$\frac{d}{ds} T(t-s)f(s, u^*(s), v^i) = T(t-s) \frac{d}{ds} f(s, u^*(s), v^i) - T(t-s)Af(s, u^*(s), v^i) \quad (58)$$

for all  $i = 1, \dots, N$ . By taking the norm in (58), applying the triangular inequality and using that  $\{T(t)\}_{t \geq 0}$  is contractive, i. e.,  $\|T(t)\|_{L(X)} \leq 1$  for all  $t \geq 0$ , we have

$$\left\| \frac{d}{ds} T(t-s)f(s, u^*(s), v^i) \right\|_X \leq \left\| \frac{d}{ds} f(s, u^*(s), v^i) \right\|_X + \|Af(s, u^*(s), v^i)\|_X. \quad (59)$$

Thus hypothesis (H<sub>2</sub>) holds, because

$$\left\| \frac{d}{ds} f(s, u^*(s), v^i) \right\|_X = \left\| \frac{d}{ds} f_i u^*(s) \right\|_X \leq \|f_i\|_X \left| \frac{d}{ds} u^*(s) \right| < \infty \quad (60)$$

and

$$\|Af(t, u^*(t), v^i)\|_X \leq \|Af_i\|_X |u^*(t)| < \infty \quad (61)$$

due to assumption (57). Also, observe that

$$\sup_{t \in (0, t_f)} \|f(t, u^*(t), v^i)\|_X \leq \|f_i\|_X \|u^*(t)\|_\infty < \infty, \quad (62)$$

because  $u^*$  is at least piecewise continuous by assumption.

Thus, by Remark 2, we can conclude that Algorithm 1 terminates in a finite number of steps with a feasible mixed-integer solution  $[u^*, v^*]$  satisfying the estimate  $|J(u^*, v^*) - J(u^*, \alpha^*)| \leq \varepsilon$ . Moreover, by Theorem 1, if  $[u^*, \alpha^*]$  is the optimal solution of the relaxed and convexified problem (9), then the integer-gap of  $[u^*, \alpha^*]$  and the optimal mixed-integer solution  $[u^*, v^*]$  is less or equal than  $\varepsilon$ . The desired switching structure  $\sigma: [0, t_f] \rightarrow \{1, \dots, N\}$  is finally given by  $\sigma(t) = v^*(t)$ .

*Example 1.* To demonstrate the applicability of the approach, we implemented Algorithm 1 for a test problem of the form (53)–(54) with a two-dimensional rectangular domain  $\Omega$  and the following parameters.

Let  $\Omega = [0, L_\xi] \times [0, L_\zeta]$ ,  $L_\xi = 1$ ,  $L_\zeta = 2$  and  $t_f = 15$  and suppose that there are given 9 actuator locations  $x_i$  with the positions given by  $(\xi_j, \zeta_k) \in \Omega$ , where

$$\xi_j = \frac{j + 0.005L_\xi}{4}, \quad \zeta_k = \frac{k + 0.005L_\zeta}{4}, \quad j, k = 1, 2, 3. \quad (63)$$

Further suppose that there is a point actuator for each of these locations  $x_i$  which we model here by setting  $f_i = B_i$  with

$$B_i(x) = \frac{1}{\sqrt{2\pi\epsilon}} e^{-\frac{(x_i - x)^2}{2\epsilon}} \quad (64)$$

for some small, but fixed  $\epsilon > 0$ . Note that

$$\int_{\Omega} B_i(x) dx = 1 \quad (65)$$

and that  $B_i(x)$  converges to the Dirac delta function  $\delta(x - x_i)$  as  $\epsilon \rightarrow 0$ .

As initial data we take

$$y_0(\xi, \zeta) = 10 \sin(\pi\xi) 10 \sin(\pi\zeta) \quad (66)$$

and as parameters in the cost function we take  $\lambda_1 = 2$  and  $\lambda_2 = \frac{1}{500}$ .

We have chosen these numerical values to match as closely as possible the two-dimensional example in [10] motivated by thermal manufacturing. The only difference is that the pointwise actuators  $\delta(x - x_i)$  were approximated in [10] by indicator functions of an epsilon environment while we choose here a smoother approximation in view of (57). Regarding a direct treatment of  $\delta(x - x_i)$  as an unbounded control operator instead of using the bounded approximation (64), see the comments in Section 6.

The solution of the relaxed optimal control problem (9) has been computed numerically. We discretized the state equation (53) in space using a standard Galerkin approach with triangular elements and linear Ansatz-functions. We eliminated one control by setting  $\tilde{\alpha}_i(t) = \alpha_i(t)$ ,  $i = 1, \dots, N-1$ , where we then get  $\alpha_N(t) = 1 - \sum_{i=1}^{N-1} \tilde{\alpha}_i(t)$  using the constraint

$$\sum_{i=1}^N \alpha_i(t) = 1, \quad t \in [0, t_f]. \quad (67)$$

This constraint is then always fulfilled and the condition

$$\alpha_N(t) \in [0, 1], \quad t \geq t_0 \quad (68)$$

is equivalent to imposing that  $\tilde{\alpha}_i \in [0, 1]$  and

$$\sum_{i=1}^{N-1} \tilde{\alpha}_i - 1 \leq 0. \quad (69)$$

The resulting semi-discretized control problem was solved with Bock's direct multiple shooting method [2, 13] implemented in the software-package MUSCOD-II. The control functions  $u$  and  $\tilde{\alpha}$  were chosen as piecewise constant and initialized with  $u(t) = 0$  and  $\tilde{\alpha}_i(t) = \frac{1}{9}$ ,  $t \in [0, t_f]$ ,  $i = 1, \dots, 8$ . Assumption (56) thus clearly holds.

The computations were made for an unstructured grid with 162 triangular elements and 8, 16 and 32 equidistant shooting intervals. Time integration was carried out by a BDF-method and sensitivities were computed using internal numerical differentiation.

$\Delta t_{\max}$	Rel. Cost $J^* = J(u^*, \tilde{\alpha}^*)$	Mix.-Int. Cost $J = J(u^*, v^*)$	Error $\frac{1}{J^*} J^* - J $
1.8750	5.634024E+04	1.283813E+05	2.9809
0.9375	4.190360E+04	7.080185E+04	1.1955
0.4688	3.224914E+04	6.175488E+04	0.9149

TABLE 1. Performance of the relaxation method for Example 1.

We implemented Algorithm 1 with adaptively solving the relaxed problem on a common control discretization grid  $\mathcal{G}^k$  for  $u$  and  $\tilde{\alpha}$  in accordance with Remark 3. For the computations, we used bisection for refinements of the control grids  $\mathcal{G}^k$  in step 7.

The performance of the relaxation method is summarized in Table 1. We see that the relative error of the mixed-integer solution compared with the best found relaxed solution decreases with the grid refinements in accordance with Remark 1. The best found controls and the evolution of the state norm of the corresponding solution are displayed in Figure 1. We see the rounding error in form of an overshooting behavior when comparing the evolution of the  $L^2(\Omega)$ -norm of the relaxed and the mixed-integer solution. This effect decreases with the size of the time discretization step size. The cost corresponding to the best found solution is 6175. Unfortunately, [10] does not report the cost of the best found solution, but a cumulative  $L^2(0, 15; L^2(\Omega))$ -norm of 90.27. The cumulative  $L^2(0, 15; L^2(\Omega))$ -norm of our best found solution is 78.58.

**5.2. A semilinear reaction-diffusion system.** Let  $\Omega$  be a bounded domain in  $\mathbb{R}^n$  with a smooth boundary  $\Gamma$  and consider the classical Lotka-Volterra system with diffusion

$$\begin{cases} \frac{\partial y_1}{\partial t}(x, t) - d_1 \sum_{j=1}^n \frac{\partial^2 y_1}{\partial x_j^2}(x, t) = y_1(x, t)(a_1 - b_1 v(t) - c_1 y_2(x, t)) & \text{in } Q \\ \frac{\partial y_2}{\partial t}(x, t) - d_2 \sum_{j=1}^n \frac{\partial^2 y_2}{\partial x_j^2}(x, t) = y_2(x, t)(a_2 - b_2 v(t) - c_2 y_1(x, t)) & \text{in } Q \\ \frac{\partial y_1}{\partial \nu}(x, t) = \frac{\partial y_2}{\partial \nu}(x, t) = 0 & \text{on } \Sigma \\ y_1(x, 0) = y_{1,0}(x), \quad y_2(x, 0) = y_{2,0}(x) & \text{in } \Omega \end{cases} \quad (70)$$

with constants  $a_i, b_i, c_i, d_i > 0$ ,  $i = 1, 2$ , domains  $Q = \Omega \times [0, t_f]$ ,  $\Sigma = \Gamma \times [0, t_f]$  and control  $0 \leq v(t) \leq 1$ . System (70) describes the interaction of two populations  $y_1$  and  $y_2$ , both spatially distributed and diffusing in  $\Omega$ . The initial distribution  $y_{1,0}, y_{2,0}$  at  $t = 0$  is assumed to be non-negative. The boundary conditions then imply that the populations  $y_1$  and  $y_2$  are confined in  $\Omega$  for all  $t \geq 0$ . The function  $v$  models a control of the system and we shall investigate to approximate optimal controls  $v^*(t)$  taking values in  $\{0, 1\}$  as to minimize the distance of the population  $(y_1, y_2)$  to its uncontrolled ( $v = 0$ ) steady state distribution  $(\bar{y}_1, \bar{y}_2)$  given by the constant functions

$$\bar{y}_1(x) = \frac{a_2}{c_2}, \quad \bar{y}_2(x) = \frac{a_1}{c_1}, \quad x \in \Omega.$$

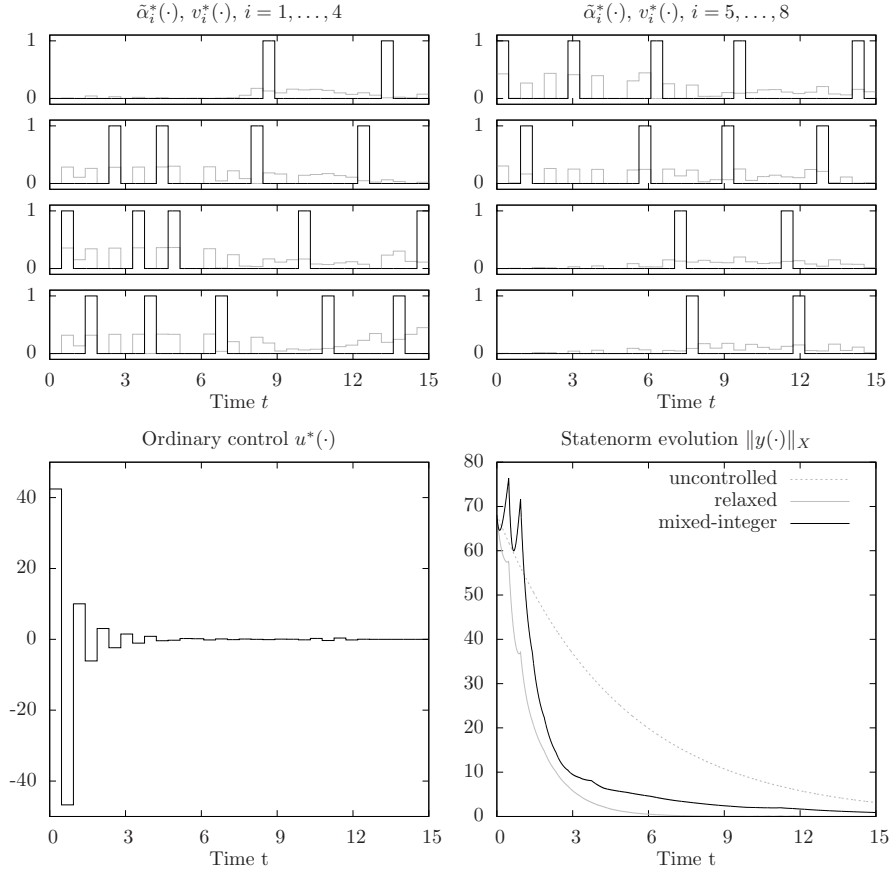


FIGURE 1. Numerical results for Example 1. The upper figures show the best found integer controls  $v_i^*(\cdot)$  and their relaxation  $\tilde{\alpha}_i^*(\cdot)$ , from bottom to top,  $i = 1, \dots, 4$  (left) and  $i = 5, \dots, 8$  (right). Control  $v_9^*(\cdot)$  is defined by  $v_9^*(t) = 1 - \sum_{i=1}^8 v_i^*(t)$ ,  $t \in [0, 15]$ . The lower figures show the corresponding ordinary control  $u^*(\cdot)$  (left) and the evolution of the state norm (right).

In order to bring the system into abstract form (2), set  $X = L^2(\Omega) \times L^2(\Omega)$ ,  $U = U_{\text{ad}} = \{0, 1\}$ ,  $V = \mathbb{R}$ ,  $V_{\text{ad}} = \{0, 1\}$ , define the operator  $A: D(A) \rightarrow X$  by

$$D(A) = \{(y_1, y_2) \in H^2(\Omega) \times H^2(\Omega) : \frac{\partial y_1}{\partial \nu}(x, t) = \frac{\partial y_2}{\partial \nu}(x, t) = 0, \text{ on } \Gamma\}$$

$$A(y_1, y_2)(x) = (d_1 \sum_{j=1}^n \frac{\partial^2 y_1}{\partial x_j^2}(x), d_2 \sum_{j=1}^n \frac{\partial^2 y_2}{\partial x_j^2}(x)), \quad (y_1, y_2) \in D(A),$$

define the non-linear function  $f: X \times U \times V = X \times V \rightarrow X$  by

$$f((y_1, y_2), v)(x) = (y_1(x)(a_1 - b_1 v - c_1 y_2(x)), y_2(x)(a_2 - b_2 v - c_2 y_2(x)))$$

and define the cost functions  $\Phi$  and  $L$  by

$$\Phi((y_1, y_2)) = 0, \quad L((y_1, y_2)) = \int_{\Omega} \|y_1(x) - \bar{y}_1(x)\|^2 + \|y_2(x) - \bar{y}_2(x)\|^2 dx.$$

It is well-known, that  $(A, D(A))$  is the generator of an analytic semigroup on  $X$  and that for any non-negative initial data  $y_{1,0}, y_{1,0} \in X$ , the system (70) has a non-negative unique mild solution  $(y_1, y_2) \in C([0, t_f], X \times X)$  for every  $v \in$



$\Delta t_{\max}$	Rel. Cost $J^* = J(u^*, \alpha^*)$	Integer Cost $J = J(u^*, v^*)$	Error $\frac{1}{J^*} J^* - J $
2.0000	7.066392E+01	8.287875E+01	0.4065
1.0000	5.978818E+01	6.958250E+01	0.1809
0.5000	5.892414E+01	5.875641E+01	0.0028

TABLE 2. Performance of the relaxation method for Example 2.

$L^\infty(0, t_f; [0, 1])$ . Moreover, for initial data  $y_{1,0}, y_{1,0} \in D(A)$ , this solution is classical and satisfies  $(y_1, y_2) \in C^1([0, t_f], D(A) \times D(A))$  for every  $v \in C_{\text{pw}}^{0,\vartheta}(0, t_f; [0, 1])$ . Existence (local in time) and uniqueness follows from classical theory for semilinear parabolic equations, see, e.g., [15, Chapter 6]. Global existence results for (70) are obtained by a-priori bounds on the solution using contracting rectangles [3].

Assume that the initial data satisfies  $y_{1,0}, y_{1,0} \in D(A)$  and that, for some  $0 < \vartheta \leq 1$ ,

$$\alpha^* \in C_{\text{pw}}^{0,\vartheta}(0, t_f; [0, 1]) \quad (71)$$

is a feasible solution for the relaxed problem (9).

We can then see that hypothesis (H<sub>1</sub>) of Theorem 1 holds, because  $\phi$  and  $L$  are continuous and because it follows from the above well-posedness results that the sets  $\mathcal{Y}_1$  and  $\mathcal{Y}_2$  are bounded.

Moreover, we claim that hypothesis (H<sub>2</sub>) of Theorem 1 holds: By analyticity of  $\{T(t)\}_{t \geq 0}$ , we have that  $s$  almost everywhere in  $(0, t_f)$ ,

$$\frac{d}{ds}T(t-s)f(y(s), v^i) = -AT(t-s)f(y(s), v^i) + T(t-s)f_y(y(s), v^i)y_s(s), \quad (72)$$

where  $f_y = \frac{d}{dy}f$  and  $y_s = \frac{d}{ds}y$ . Using that  $y(s) \in D(A) \times D(A)$  for all  $s \in [0, t_f]$  and  $f: D(A) \rightarrow D(A)$ , we see that

$$\| -AT(t-s)f(y(s), v^i) \|_X \leq \|T(t-s)\|_{\mathcal{L}(X)} \|Af(y(s), v^i)\|_X < \infty. \quad (73)$$

Using that  $f$  is a smooth function, we see that

$$\|T(t-s)f_y(y(s), v^i)y_s(s)\|_X \leq \|T(t-s)\|_{\mathcal{L}(X)} \|f_y(y(s), v^i)\|_X \|y_s(s)\|_X < \infty. \quad (74)$$

Thus (72) yields the estimate

$$\left\| \frac{d}{ds}T(t-s)f(y(s), v^i) \right\|_X < \infty \quad (75)$$

for  $s \in [0, t_f]$  a. e.

By well-posedness of the problem (70) for every  $v \in C_{\text{pw}}^{0,\vartheta}(0, t_f; [0, 1])$ , with  $\vartheta$  from assumption (71) we get the estimate

$$\sup_{t \in [0, t_f]} \|f(t, y^*(t), v^i)\|_X < \infty, \quad (76)$$

verifying hypothesis (H<sub>3</sub>).

Altogether, by Remark 2, we can conclude that Algorithm 1 terminates in a finite number of steps with a feasible integer solution  $v^*$  satisfying the estimate  $|J(v^*) - J(\alpha^*)| \leq \varepsilon$ . Again, by Theorem 1, if  $\alpha^*$  is the optimal solution of the relaxed and convexified problem (9), then the integer-gap of  $\alpha^*$  and the optimal integer solution  $v^*$  is less or equal than  $\varepsilon$ .

*Example 2.* We applied Algorithm 1 to a semilinear test problem of the form (70) again for a two dimensional domain  $\Omega$  with the following parameters.

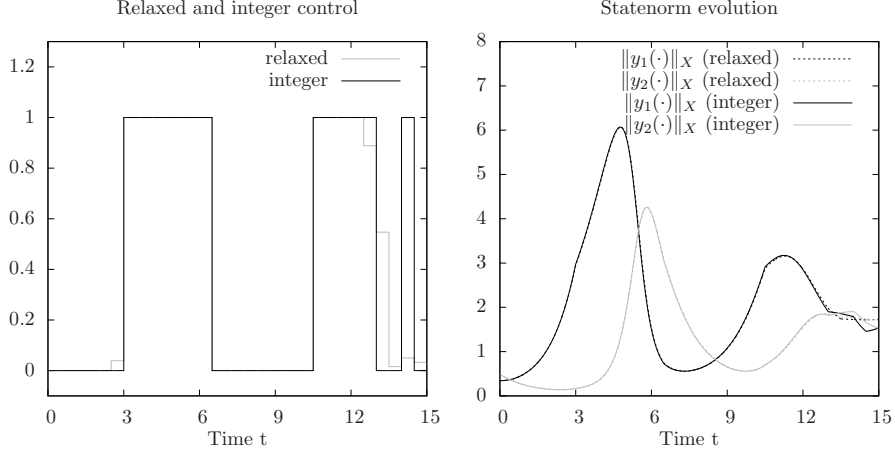


FIGURE 2. Numerical results for Example 2. The left figure shows the best found relaxed and integer control  $\alpha^*(\cdot)$ ,  $v^*(\cdot)$  and the right figure shows the corresponding evolutions of the populations  $y_1(\cdot)$ ,  $y_2(\cdot)$ .

Let  $\Omega$  being a circle with radius 1 centered at  $(1, 1)$  and choose  $a_1 = a_2 = c_1 = c_2 = 1$ ,  $b_1 = \frac{7}{10}$ ,  $b_2 = \frac{1}{2}$ , initial data  $y_{1,0}, y_{2,0} \in D(A)$  approximated by  $\tilde{y}_{1,0}(x) = \frac{1}{2}d_{\frac{1}{2}}(x-1)$ ,  $\tilde{y}_{2,0}(x) = \frac{7}{10}d_{\frac{1}{2}}(x-1)$ , where  $d_\epsilon(x)$  given by

$$d_\epsilon(x) = \frac{1}{\sqrt{2\pi\epsilon}} e^{-\frac{x^2}{2\epsilon}} \quad (77)$$

models a population concentrated at the origin. For  $v(t) = 0$ ,  $t \geq 0$ , the solution  $y_1(t, x)$ ,  $y_2(t, x)$  converges asymptotically to a spatially constant and temporarily non-constant, periodic solution.

The computations for the optimal control are made by the same numerical method as in the previous example, but using a grid with 258 finite elements. The choice of piecewise constant controls ensures that (71) holds with  $\vartheta = 1$ . For the performance of the relaxation method implementing Algorithm 1 see Table 2. The best found controls and the evolution of the state norm of the corresponding solutions are displayed in Figure 2. Again we see a decrease of the integer-approximation error in accordance with Remark 1. The best found integer control yields a cost of 58.76.

## 6. CONCLUSIONS AND OPEN PROBLEMS

We considered mixed-integer optimal control problems for abstract semilinear evolution equations and obtained conditions guaranteeing that the solution of a relaxed optimal control problem can be approximated with arbitrary precision using a control that satisfies integer restrictions. In particular, our approach is constructive and gives rise to a numerical method for mixed-integer optimal control problems with certain partial differential equations. Moreover, we showed how these conditions imply a-priori estimates on the quality of the solution when combinatorial constraints are enforced.

Compared to the results on mixed-integer optimal control problems with ordinary differential equations [18, 16], the setting treated in this paper involves a differential operator  $A$ , taken to be a generator of a strongly continuous semigroup. This requires careful regularity considerations. When  $A$  is a Laplace operator, we showed on a linear and a semilinear example how such regularity assumptions can

be met and provided numerical examples demonstrating the practicability of the approach.

It is clear that the methodology considered in this paper generalizes to the case when the generator  $A$  of a strongly continuous semigroup is replaced by a family  $\{A(t)\}_{t \in [0, t_f]}$  of unbounded linear operators generating an evolution operator in the sense of [12]. On the other hand it is not so clear how to extend the results in case of unbounded control action, for example, Neumann or Dirichlet boundary control for the heat equation. Recalling the density of solutions to (5) in the set of solution to (6) which motivated our approach, we note that the case of unbounded control is neither covered by the available results on operator differential inclusions. While in principle semigroup techniques can deal with unbounded control operators, see for example the exposition in [1, Chapter 3], this extension is non-trivial and requires additional work.

#### ACKNOWLEDGEMENTS

Financial support of the Mathematics Center Heidelberg (MATCH), of the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences, and of the EU project EMBOCON under grant FP7-ICT-2009-4 248940 is gratefully acknowledged.

#### REFERENCES

- [1] Alain Bensoussan, Giuseppe Da Prato, Michel C. Delfour, and Sanjoy K. Mitter. *Representation and control of infinite-dimensional systems. Vol. 1.* Systems & Control: Foundations & Applications. Birkhäuser Boston Inc., Boston, MA, 1992.
- [2] H.G. Bock and K.J. Plitt. A Multiple Shooting algorithm for direct solution of optimal control problems. In *Proceedings of the 9th IFAC World Congress*, pages 242–247, Budapest, 1984. Pergamon Press.
- [3] Peter N. Brown. Decay to uniform states in ecological interactions. *SIAM J. Appl. Math.*, 38(1):22–37, 1980.
- [4] F. S. de Blasi and G. Pianigiani. Evolution inclusions in non-separable Banach spaces. *Comment. Math. Univ. Carolin.*, 40(2):227–250, 1999.
- [5] S. Engell and A. Toumi. Optimisation and control of chromatography. *Computers and Chemical Engineering*, 29:1243–1252, 2005.
- [6] Halina Frankowska. A priori estimates for operational differential inclusions. *J. Differential Equations*, 84(1):100–128, 1990.
- [7] Martin Gugat. Optimal switching boundary control of a string to rest in finite time. *ZAMM Z. Angew. Math. Mech.*, 88(4):283–305, 2008.
- [8] Falk M. Hante and Günter Leugering. Optimal boundary control of convection-reaction transport systems with binary control functions. In *Hybrid systems: computation and control*, volume 5469 of *Lecture Notes in Comput. Sci.*, pages 209–222. Springer, Berlin, 2009.
- [9] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, New York, 2009.
- [10] Orest V. Iftime and Michael A. Demetriou. Optimal control of switched distributed parameter systems with spatially scheduled actuators. *Automatica J. IFAC*, 45(2):312–323, 2009.
- [11] Y. Kawajiri and L.T. Biegler. A nonlinear programming superstructure for optimal dynamic operations of simulated moving bed processes. *IEEC Research*, 45(25):8503–8513, 2006.
- [12] S. G. Kreĭn. *Linear differential equations in Banach space*. American Mathematical Society, Providence, R.I., 1971. Translated from the Russian by J. M. Danskin, Translations of Mathematical Monographs, Vol. 29.
- [13] D.B. Leineweber, I. Bauer, A.A.S. Schäfer, H.G. Bock, and J.P. Schlöder. An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization (Parts I and II). *Computers and Chemical Engineering*, 27:157–174, 2003.
- [14] Xun Jing Li and Jiong Min Yong. *Optimal control theory for infinite-dimensional systems*. Systems & Control: Foundations & Applications. Birkhäuser Boston Inc., Boston, MA, 1995.
- [15] A. Pazy. *Semigroups Of Linear Operators And Applications To Partial Differential Equations*. Applied Mathematical Sciences Series, Springer-Verlag, New York, 1983.
- [16] S. Sager, H.G. Bock, and M. Diehl. The integer approximation error in mixed-integer optimal control. *Mathematical Programming*, 2010. DOI 10.1007/s10107-010-0405-3.

- [17] Sebastian Sager. Reformulations and algorithms for the optimization of switching decisions in nonlinear optimal control. *Journal of Process Control*, 19(8):1238–1247, 2009.
- [18] Sebastian Sager, Hans Georg Bock, and Gerhard Reinelt. Direct methods with maximal lower bound for mixed-integer optimal control problems. *Math. Program.*, 118(1, Ser. A):109–149, 2009.
- [19] Sebastian Sager, Michael Jung, and Christian Kirches. Combinatorial integral approximation. *Math. Methods Oper. Res.*, 73(3):363–380, 2011.
- [20] Tomoyo Sakata, David K. Jackson, Shu Mao, and Gerard Marriott. Optically switchable chelates: Optical control and sensing of metal ions. *J Org Chem*, 73(1):227–233, 2008.
- [21] Fredi Tröltzsch. *Optimal control of partial differential equations*, volume 112 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.
- [22] M. van Sint Annaland, J.A.M. Kuipers, and W.P.M. van Swaaij. Safety analysis of switching between reductive and oxidative conditions in a reaction coupling reverse flow reactor. *Chemical Engineering Science*, 56(4):1517 – 1524, 2001.
- [23] Enrique Zuazua. Switching control. *J. Eur. Math. Soc.*, 13:85–117, 2011.